The Causal Effect of Stay-At-Home Orders: a synthetic control study in the San Francisco Bay Area

Mehrdad Yazdani
myazdani@gmail.com

1 Overview

The COVID-19 outbreak in December 2019 is an unprecedented modern global pandemic. With no vaccination in sight, policy makers are left with transmission and contact reduction interventions to slow the virus spread [6]. However, many of these measures accrue enormous economic costs. Stay-at-home orders and business closures in particular have severe economic consequences forcing policy makers to decide between "flattening the curve" [10] or re-mobilizing the economy [11]. As a result, it is critical to understand the effectiveness of such measures.

A common, but limited and misleading, approach of understanding the effectiveness of a reduction strategy is to compare regions with and without reduction measures. The fallacy in drawing conclusions from such comparisons is that the comparison does not constitute a properly randomized control trial. Although running randomized control trials are impossible in this case, the method of "synthetic control" provides a solution to answer counterfactual "what if" questions [3, 2].

To understand the effectiveness of stay-at-home orders, we consider a case study in California. On March 16, 2020 six counties in the San Francisco Bay Area became the first regions in the United States to place stay-at-home orders (referred to as "shelter in place" or SIP). On March 20, 2020 California became the first state in the US to issue SIP orders to the entire state.

In this paper we investigate if the Bay Area *had not* started SIP 4 days earlier, would there have been a significant change in number of COVID-19 cases? By creating a synthetic Bay Area, we are able to answer this question and we observe that the early SIP order had a significant impact in reducing the number of cases by as much as 10 times. As a result, we can conclude that even though the Bay Area initiated SIP by just 4 days, the intervention contributed significantly to flattening the curve.

2 Data and Methodology

We use data provided by Johns Hopkins University [8, 1]. The six San Francisco Bay Area counties that had early SIP intervention on March 16 are: Alameda, Contra Costa, Marin, San Francisco, San Mateo, and Santa Clara. We group these regions together and refer to them as the Bay Area or the treatment group. The remaining counties in California form our donor pool that we use to create a synthetic Bay Area [3, 2]. We only consider the number of COVID-19 cases and ignore the number of deaths. We also ignore counties from the donor pool that reported less than 5 COVID-19 cases on the intervention date of March 16, 2020. This leaves us with 10 donor counties: Los Angeles, Orange, Placer, Riverside, Sacramento, San Diego, San Joaquin, Santa Cruz, Solano, and Sonoma.

Unlike [3, 2], we do not constrain the formation of the synthetic control arm to be a convex combination of the donor pool (as similarly relaxed in [5, 4]). A particular challenge with COVID-19 data is the number of cases are under-reported. The robust synthetic control method [5] mitigates this issue by solving a matrix completion problem with Singular Value Decomposition (SVD). However, since our count data is strictly positive, a further constrained matrix factorization method such as Non-negative Matrix Factorization (NMF) [9, 7] may yield better regularization and generalization. We compare SVD and NMF by comparing their errors on the validation set. Our validation set is the

Preprint. Under review.

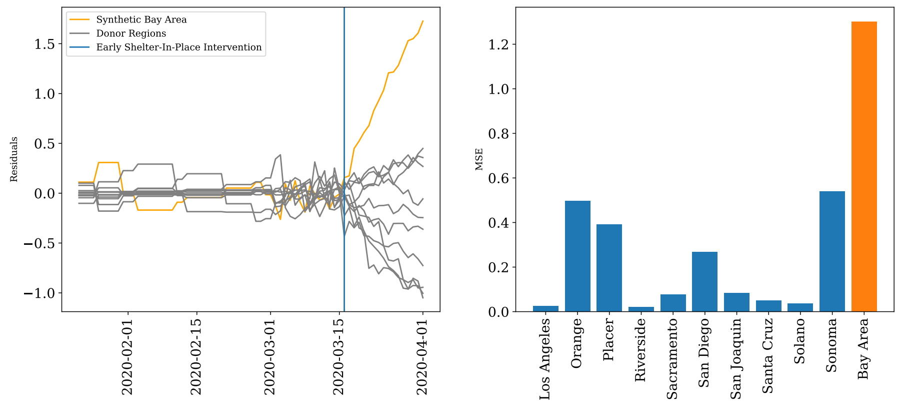

Figure 2: Residuals are predictions minus targets. MSE is computed after the intervention.

data from March 13-16, our training set is the available data prior to March 13, and we evaluate on all data after March 16 (the intervention date) until April 1, 2020. We do not consider evaluating on data after April 1 since the out of sample errors are increasingly large. All data is log-transformed[1].

3 Results

We validate the hyper-parameters and model selection with the validation data as described in section 2. We find that NMF generalizes better and validates the hypothesis that a constrained matrix factorization model has better properties than using just SVD as done in [5]. Figure 1 shows the number of COVID-19 cases in the Bay Area as observed and compares against the Synthetic Bay Area. Synthetic Bay Area is similar to the actual Bay Area before the onset of the intervention on March 16. This suggests that we can use Synthetic Bay Area as our synthetic control. As we extrapolate beyond the intervention, we see that the number of COVID-19 cases for Synthetic Bay Area increase significantly (by over 10 times).

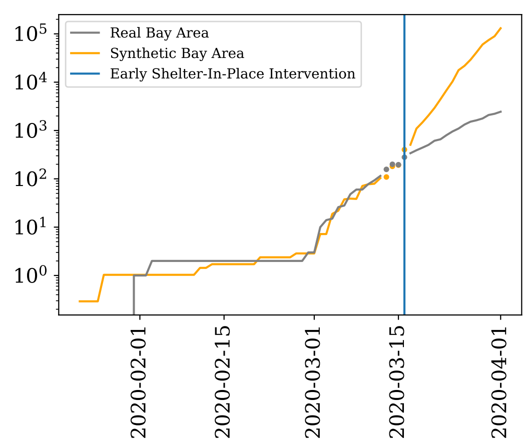

Figure 1: Number of COVID-19 cases

To ensure that our result is not a false positive, we also run placebo comparisons with the donor counties as done in [3, 2]: for each of the donor counties we use the same methodology for creating synthetic placebo controls and measure their residuals. Figure 2 shows residuals of the donor counties in the placebo studies. As was in the case for Synthetic Bay Area, we are able to successfully create a synthetic control for each of the placebo counties. We see that the residuals for the placebos are significantly smaller than Synthetic Bay Area, suggesting that the early SIP intervention had a significant impact for the Bay Area.

4 Conclusions

The early onset of a pandemic is particularly prone to underestimates in exponential growth rates. Our study suggests that decisions delayed by even a few days can have enormous impact in the spread of disease. Synthetic control is an invaluable tool in addressing counterfactual questions as posed here. However our analysis indicates care must be taken to incorporate prior knowledge into the latent representation of the data when feasible.

[1] Source code and additional figures: `https://github.com/myazdani/shelter-in-place`

References

[1] Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university. `https://github.com/CSSEGISandData/COVID-19`, 2020.

[2] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[3] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

[4] Muhammad Amjad, Vishal Misra, Devavrat Shah, and Dennis Shen. mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–27, 2019.

[5] Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

[6] Albert-László Barabási et al. *Network Science: Chapter 10*. Cambridge university press, 2016.

[7] Andrzej Cichocki and Anh-Huy Phan. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 92(3):708–721, 2009.

[8] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 2020.

[9] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[10] Richard J Hatchett, Carter E Mecher, and Marc Lipsitch. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proceedings of the National Academy of Sciences*, 104(18):7582–7587, 2007.

[11] Linda Thunström, Stephen C Newbold, David Finnoff, Madison Ashworth, and Jason F Shogren. The benefits and costs of using social distancing to flatten the curve for covid-19. *Journal of Benefit-Cost Analysis*, pages 1–27, 2020.